Published Online First on February 10, 2009

Perspective

Rembrandt: Helping Personalized Medicine Become a Reality through Integrative Translational Research

Subha Madhavan,[1] Jean-Claude Zenklusen,[2] Yuri Kotliarov,[2] Himanso Sahni,[3] Howard A. Fine,[2] and Kenneth Buetow[1]

[1]Center for Biomedical Informatics and Information Technology and [2]Center for Cancer Research, Neuro-Oncology Branch, National Cancer Institute, Bethesda, Maryland; and [3]Science Applications International Corporation, San Diego, California

Abstract

Finding better therapies for the treatment of brain tumors is hampered by the lack of consistently obtained molecular data in a large sample set and the ability to integrate biomedical data from disparate sources enabling translation of therapies from bench to bedside. Hence, a critical factor in the advancement of biomedical research and clinical translation is the ease with which data can be integrated, redistributed, and analyzed both within and across functional domains. Novel biomedical informatics infrastructure and tools are essential for developing individualized patient treatment based on the specific genomic signatures in each patient's tumor. Here, we present Repository of Molecular Brain Neoplasia Data (Rembrandt), a cancer clinical genomics database and a Web-based data mining and analysis platform aimed at facilitating discovery by connecting the dots between clinical information and genomic characterization data. To date, Rembrandt contains data generated through the Glioma Molecular Diagnostic Initiative from 874 glioma specimens comprising ~566 gene expression arrays, 834 copy number arrays, and 13,472 clinical phenotype data points. Data can be queried and visualized for a selected gene across all data platforms or for multiple genes in a selected platform. Additionally, gene sets can be limited to clinically important annotations including secreted, kinase, membrane, and known gene-anomaly pairs to facilitate the discovery of novel biomarkers and therapeutic targets. We believe that Rembrandt represents a prototype of how high-throughput genomic and clinical data can be integrated in a way that will allow expeditious and efficient translation of laboratory discoveries to the clinic.

(Mol Cancer Res 2009;7(2):157–67)

Introduction

Primary brain tumors are a leading cause of cancer mortality in adults and children in the United States (1). The molecular and genetic heterogeneity of gliomas undoubtedly contributes to the varied and often suboptimal response to treatment that is usually predicated on standard pathologic diagnoses. Improvement in the prognosis of patients with gliomas will likely come about through the use of new targeted therapies based on the biological knowledge of the tumors at a molecular level.

To identify glioma-specific targets, consistent molecular characterization of a large number of tumors is required. To date, all the studies published have limitations due to incomplete coverage of whole-genome expression due to the usage of small or outdated, legacy, microarray platforms (2, 3), limited number of samples studied and/or incomplete inclusion of various different glioma subtypes and grades (4, 5), or the narrow scope of targets being investigated. Thus, we have put together a national, publicly funded effort that we call the Glioma Molecular Diagnostic Initiative (GMDI), which, coupled with its bioinformatics counterpart, Repository of Molecular Brain Neoplasia Data (Rembrandt), is designed to breach the gap of biological information related to primary brain tumors to help patients receive a better, biologically oriented therapy tailored to their specific needs.

Rembrandt is a powerful and intuitive informatics system designed to integrate genetic and clinical information for improved research, disease diagnosis, and treatment (as shown in Fig. 1). The platform supports clinical genomic research and (as data are collected and analyzed) will create a knowledge base that allows physicians to predict clinical outcomes and therapeutic efficacy based on an individual's clinical and genetic profiles, thereby enabling personalized medicine.

To support discovery, the Rembrandt platform also allows researchers to search, import, and aggregate additional data from internal and external databases (such as GenBank, University of California at Santa Cruz golden path data sets, and Biocarta pathways), analyze the combined data sets to identify meaningful patterns (including specific chromosomal

Received 9/17/08; revised 11/6/08; accepted 11/10/08; published OnlineFirst 02/10/2009.

Grant support: Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research and National Institute of Neurological Disorders and Stroke.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Molecular Cancer Research Online (http://mcr.aacrjournals.org/).

S. Madhavan and J-C. Zenklusen contributed equally to this work.

Requests for reprints: Jean-Claude Zenklusen, Center for Cancer Research, Neuro-Oncology Branch, National Cancer Institute, 37 Convent Drive, Room 1142B, MSC 4254, Bethesda, MD 20892. Phone: 301-451-2144; Fax: 301-480-4743. E-mail: jz44m@nih.gov

doi:10.1158/1541-7786.MCR-08-0435

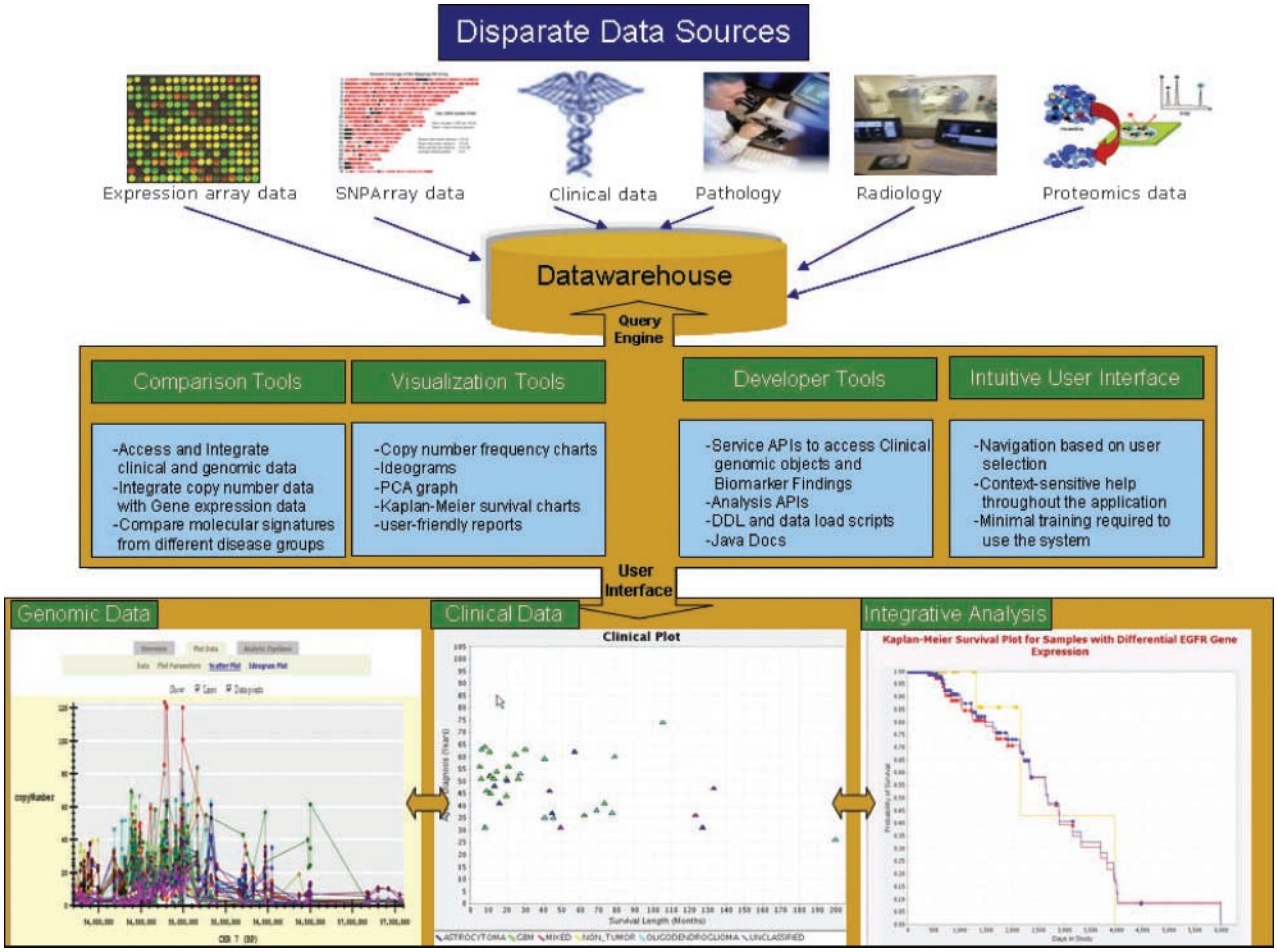


FIGURE 1. Data integration via the Rembrandt discovery platform.

abnormalities), and share their research with other physicians and researchers within their own institution or in other physical locations. Each user is assigned a specific role that governs how much of the study data are accessible. A series of intuitive tools enable users to easily analyze and interact with the integrated data to achieve greater insight into molecular signatures that characterize each tumor and correlate with clinical outcome.

Unlike many biomedical database systems, Rembrandt is a fully integrated platform that supports multiple facets of clinical and molecular research, discovery, and hypothesis generation. This shared environment crosses many disciplines including genetic research and clinical care. As such, the platform should serve to foster cooperation and integration between research and clinical disciplines and expedite the time and increase the depth to which molecular data become relevant to the clinical environment.

Materials and Methods

Glioma Molecular Diagnostic Initiative

Sample Acquisition and Diagnosis. To better understand the genetic pathogenesis of gliomas and begin to identify potential glioma-specific molecular therapeutic targets, consistent molecular characterization of a large number of tumors is required.

This process was undertaken under a national prospective clinical trial that would eventually be institutional review board approved both within the National Cancer Institute intramural program and through both Cancer Therapy Evaluation Program-sponsored adult brain tumor consortia (NABTT and NABTC protocol 01-07). With the activation of this study, we collected matched tumor, blood, and plasma from the 14 contributing institutions (NIH, Henry Ford Hospital, Thomas Jefferson University, University of California at San Francisco, H. Lee Moffitt Hospital, University of Wisconsin, University of Pittsburgh Medical Center, University of California at Los Angeles, The University of Texas M. D. Anderson Cancer Center, Dana-Farber Cancer Center, Duke University, Johns Hopkins University, Massachusetts General Hospital, and Memorial Sloan Kettering Cancer Center). All tissues collected are sent to the Neuro-Oncology Branch laboratory for processing. The samples were provided as snap-frozen sections of areas immediately adjacent to the region used for the histopathologic diagnosis. Initial histopathologic diagnosis is done at the tissue collecting institution following the WHO standards (6). The initial diagnosis is reviewed by in-house neuropathologists to assure a measure of consistency across samples. To date, 874 complete frozen sample sets have been accrued, of those 389 are glioblastoma multiformes, 122 are astrocytomas, 113 are oligodendrogliomas, and 33 are mixed, with the reminder still unclassified.

Clinical data on the patients are collected prospectively until the patient's death through the NABTC Operations Office at The University of Texas M. D. Anderson Cancer Center and the NABTT Operations office at the Johns Hopkins University. The clinical data collected are updated into the Rembrandt database on a quarterly basis.

To assure consistency in the collection, shipment, processing, assaying, storage, data retrieval, and dissemination, we have put together a series of standard operating procedures that have resulted in a streamlined, high-throughput operation capable of handling large numbers of samples in a consistent, operator-independent fashion. Consistency of data over time is continuously monitored by looking for any signs of batch effect in the analyses.

mRNA Extraction and Gene Expression Data Processing. Tissue (~50-80 mg) from each tumor was used to extract total RNA using the Trizol reagent (Invitrogen) following the manufacturer's instructions. The quality of RNA obtained was verified with the Bioanalyzer System (ref. 7; Agilent Technologies) using the RNA Pico Chips. RNA (5 μg) extracted from the accrued samples has been processed using U133 2 Plus mRNA expression chips (Affymetrix), which contains >54,000 probe sets analyzing the expression level of >47,000 transcripts and variants, including 38,500 well-characterized human genes.

All arrays were confirmed to be within an acceptable minimal quality-control according to internal standard operating procedure variables following these criteria: (*a*) A scaling factor of <5 when the expression values are scaled to a target mean signal intensity of 500. (*b*) Signal intensity ratios of the 3′ to 5′ end of the internal control genes of β-actin and GAPDH < 3. (*c*) Affymetrix spike control (BioC, BioDN, and CreX) are always present, and percentage present calls is >35% for brain tissue.

The .cel and .txt files of all the arrays that passed the minimal quality-control were input into dChip for normalization. The model-based expression index algorithm implemented in dChip selects an invariant set with a small within-subset rank difference to serve as basis for adjusting the brightness of the arrays to a comparable level. The normalization was done at the PM and MM probe levels, and model-based expression levels were calculated using normalized probe level data. We choose the average difference model (PM > MM) to compute expression values; negative average differences were truncated to 1 or log-transformed values of zeros to flag negative signal intensities with no biological meaning.

For data preprocessing, probe-level data were consolidated into probe-set data using the Affymetrix MAS5 algorithm, with the target scaling value at 500. Probe-level data were also processed with custom Chip Definition Files (1) that rearranged Affymetrix probes into gene-based probe sets. Probes mapped to alternatively spliced exons were grouped into distinct probe sets. Most 3′ probes were selected for processing. Nonspecific probes were masked before processing.

Single tumor samples were compared with the nontumor pool and the sample average to the nontumor pool. Samples were averaged based on tumor subtypes in six categories: glioblastoma multiforme, oligodendroglioma, astrocytoma, mixed, unclassified, and unknown tumors. Group comparisons were done in R with two sample *t* tests. Signal values were first transformed to logarithm (base 2). The averages of the $\log_2$ signals of tumor and nontumor groups were computed. The magnitude of the differences between the geometric means of expression levels for each reporter from the two groups was computed. The significance of the differences between tumors (or each tumor subtype) and nontumor samples for each reporter was also evaluated.

For each individual tumor sample, signals for each tumor and the ratio between each tumor and the average of normal (geometric means, computed the same way as described above)

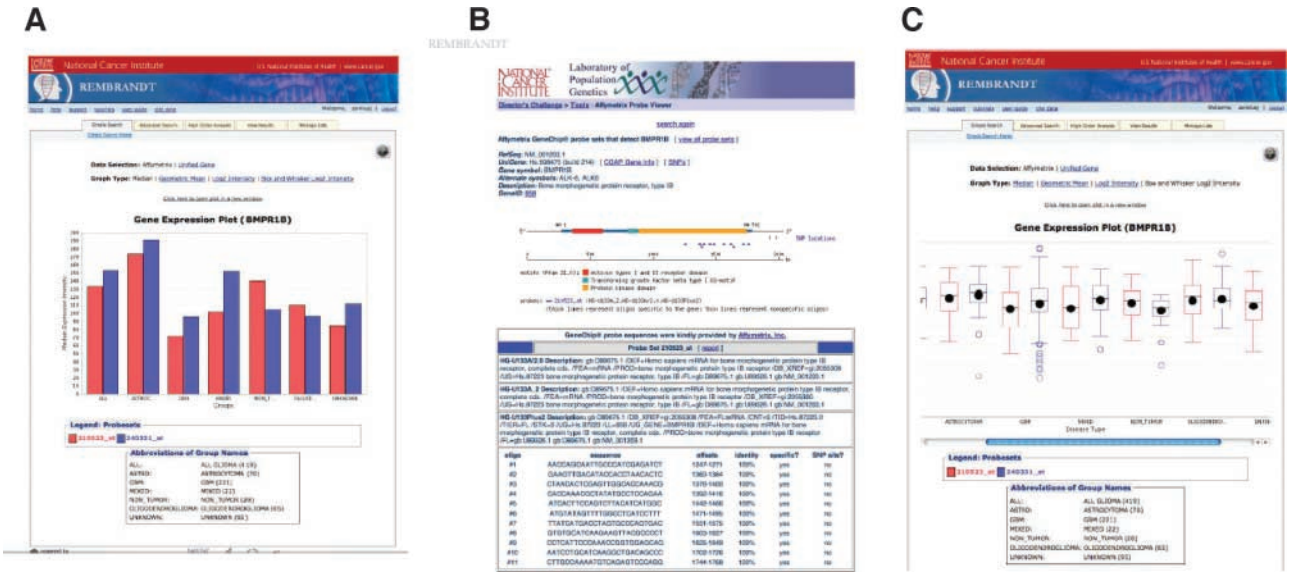
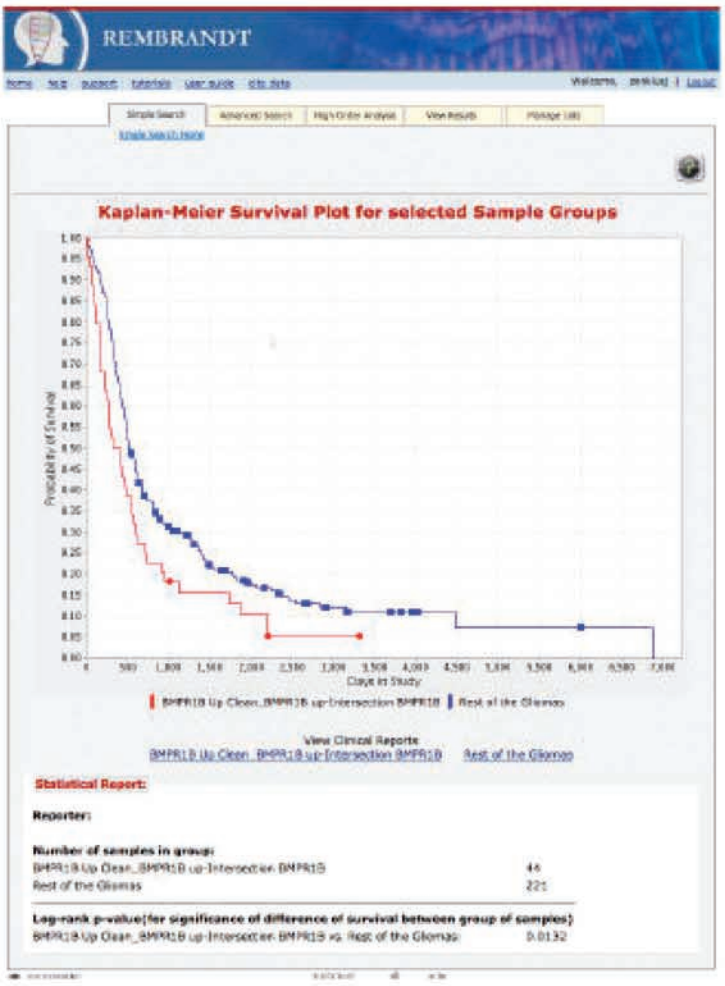
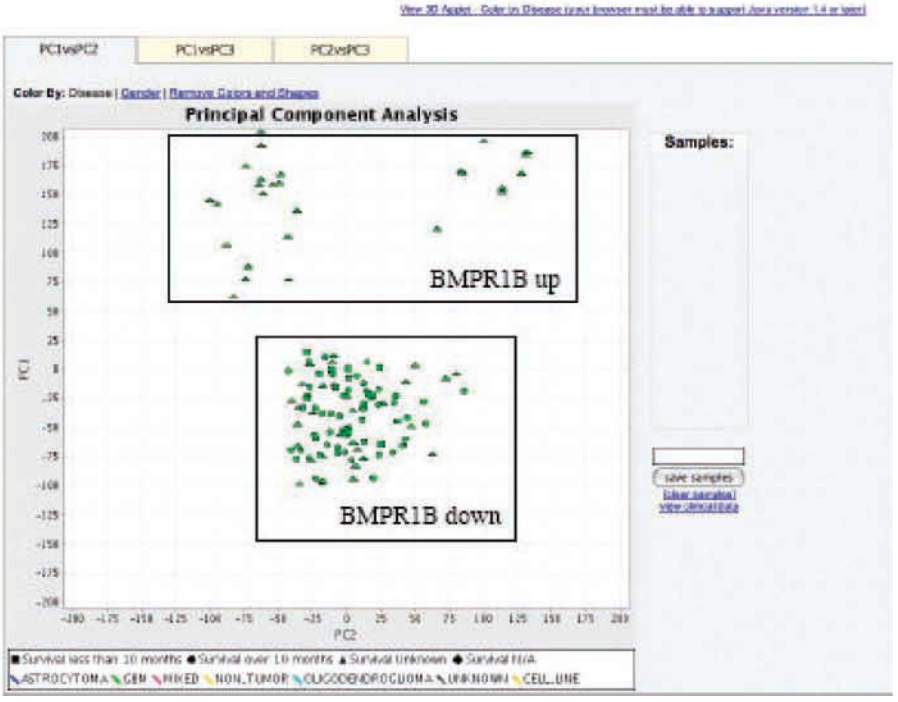


FIGURE 2. **A.** Gene expression box plot for BMPR1B. Samples are categorized by histologic type. Different Affymetrix probe sets are shown as different color bars. **B.** BMPR1B probe set in Affymetrix probe-set viewer. Information for selected probe set can be displayed, allowing the user to decide on the quality of information retrieved. **C.** BMPR1B probe set of interest showing outliers in glioblastoma multiforme samples. The ability to display expression graphs in different formats allows the use to gain a wealth of information without having to redo the queries.

A


REMBRANDT
Kaplan-Meier Survival Plot for selected Sample Groups
Probability of Survival
Days in Study
BMPR1B Up Clean_BMPR1B up-Intersection BMPR1B
Rest of the Gliomas
View Clinical Reports
BMPR1B Up Clean_BMPR1B up-Intersection BMPR1B
Rest of the Gliomas
Statistical Report:
Reporter:
Number of samples in group:
BMPR1B Up Clean_BMPR1B up-Intersection BMPR1B 44
Rest of the Gliomas 221
Log-rank p-value(for significance of difference of survival between group of samples)
BMPR1B Up Clean_BMPR1B up-Intersection BMPR1B vs. Rest of the Gliomas 0.0132

B


PC1vsPC2
PC1vsPC3
PC2vsPC3
Color By: Disease | Gender | Remove Grade and Stages
Principal Component Analysis
BMPR1B up
BMPR1B down
PC1
PC2
Samples:
save samples
clear samples
view clinical data
Survival less than 10 months Survival over 10 months Survival Unknown Survival N/A
ASTROCYTOMA GBM MIXED NON_TUMOR OLIGODENDROGLIOMA UNKNOWN CELL_LINE

were computed. All processes were done separately for various data groups (public data and institution-based data).

DNA Extraction and Genomic Alteration Analysis. Tissue (~10 μg; as recommended by the manufacturer) from each tumor was used to extract high molecular weight, genomic DNA using QIAamp DNA Micro DNA extraction kit (Qiagen) following the manufacturer's instructions. The quality of DNA was checked by electrophoresis run in a 2% agarose gel.

Genomic DNA (250 ng) from samples received has been hybridized to 100K single nucleotide polymorphism chips (ref. 8; Affymetrix), which covers 116,204 single nucleotide polymorphism loci in the human genome with a mean intermarker distance of 23.6 kb. These arrays give two simultaneous data types: allelic calls and signal intensity, allowing for the determination of both copy number alterations and regions of allelic imbalances (loss of heterozygosity). Calls were determined by the GTYPE software version 3.0 with 25% level of confidence. Only samples with call rates of >90% were accepted for any analysis.

Clinical Data Processing. The University of Texas M. D. Anderson Cancer Center serves as the operating center for clinical data collection for the GMDI trial. Clinical data reports from the case report forms were accessed through the Data Management Initiative Web portal at The University of Texas M. D. Anderson Cancer Center, parsed, and uploaded to the Rembrandt data warehouse after various preprocessing and data validations steps. The clinical data collected are updated into the Rembrandt database on a quarterly basis.

Results

A Rembrandt Storyboard

To exemplify the powerful integration that Rembrandt provides to analyze a large data set of both molecular and clinical data, we would like to show how one could come about to explore the validity of a scientific hypothesis using the system.

Suppose that one would have come across two publications on Glioma Tumor Stem Cells that mentioned the irregular expression of BMPR1B in such cells (9, 10).

A typical Rembrandt usage scenario might be to ask if BMPR1B is a potential therapeutic target as it has been recently been postulated to be involved in cell differentiation. To answer this question, a researcher can take a stepwise workflow approach in Rembrandt as shown in Figs. 2 and 3.

1. Explore the expression levels of BMPR1B in different subtypes of glioma. Analysis of the box plots in Rembrandt (Fig. 2A) indicates that probe-set 210523_at is differentially expressed in glioblastoma multiformes when compared with nontumors (borderline significance: $P < 0.04$).
2. Where does this probe map onto the transcripts of BMPR1B? Review of probe mapping in Affymetrix probe viewer integrated into Rembrandt (Fig. 2B) shows that this probe maps to coding region.
3. Are there two subpopulations of BMPR1B regulating samples? Review of the "box and whisker" plot in Fig. 2C indicates that glioblastoma multiformes have low-end outliers for BMPR1B expression.
4. Now, can we identify samples that show high (up >2) and low (down <1.5) expression of BMPR1B? Advanced queries can be set up in the Rembrandt application to create sample sets with separate up-regulation and down-regulation criteria for BMPR1 expression.
5. Does BMPR1 up-regulation affect survival? Can this sample group be compared with the rest of the gliomas? Figure 3A shows the difference in probability of survival between BMRP1 up-regulating group and the rest of the gliomas. Results indicate that BMPR1B up-regulation is bad as a prognostic factor and could be a good target for therapy.
6. How different are these sample groups beyond BMPR1B expression? By analyzing the whole gene expression patterns in both groups using the high-order analysis tool of principal component analysis (PCA; Fig. 3B), it is possible to see that BMPR1B overexpressors and underexpressors are indeed quite different at a global expression level, suggesting that this gene may hold a key to glioma diversity.

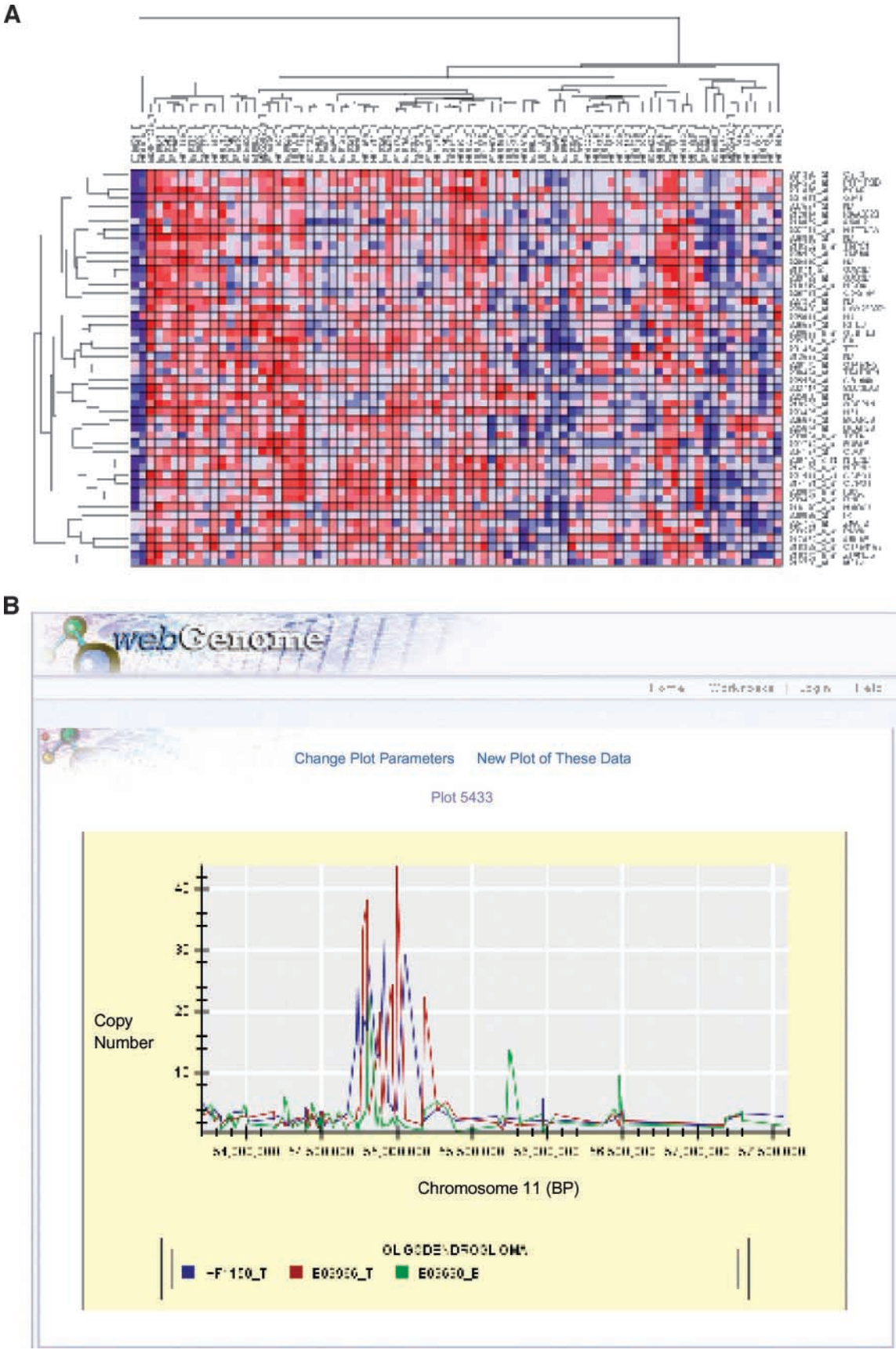
The storyboard here presented indicates that Rembrandt can effectively be used to test *in silico* a scientific hypothesis and allow for additional experimentation to occur. In fact, this has been the case with the scenario here presented and we have shown that BMPR1B is able in fact to modulate the tumorigenic potential of glioma cells (11). Additionally, a Rembrandt search of newly identified (NF1) and well-known (IGFBP2) targets of deregulation in gliomas shows that the result produced by our data set are concordant with the current knowledge of clinical features (Supplementary Fig. S1).

Key Features in Rembrandt

Integrating Genome Characterization Data with Clinical Outcomes

Users can query gene expression or copy number data and graph changes in survival rate at each time point in the study. Kaplan-Meier estimates are calculated based on the last

FIGURE 3. A. Kaplan-Meier survival plot showing survival comparing BMRP1 up-regulating samples and the rest of the gliomas in the database. This plot allows the identification of putative clinically relevant genes to explore as new targets for therapy. Users can query gene expression and graph changes in survival rate at each time point on the study. Kaplan-Meier estimates are calculated based on the last follow-up time and the censor status (0, alive; 1, dead) from the samples of interest. Kaplan-Meier estimates are then plotted against the survival time. Users can dynamically modify the fold change (up-regulation and down-regulation) thresholds and redraw the plot. A log-rank *P* value is provided as an indication of significance of the difference in survival between any two groups of samples segregated based on gene expression of the gene of interest. **B.** Performing PCA and correlating with clinical data. An example of PCA report from the Rembrandt application. These two-dimensional (*top*) and three-dimensional (*bottom*) graphs plot the various principal components from the gene expression PCA. Various analysis options are provided to the user to select from an array of gene/reporter filtering and sample selection settings. Users can select samples in the two-dimensional plot to retrieve related clinical information on the selected patients.


A
B
webGenome
Change Plot Parameters
New Plot of These Data
Plot 5433
Copy Number
Chromosome 11 (BP)

follow-up time and the censor status (0 = alive, 1 = dead) from the samples of interest. Kaplan-Meier estimates are then plotted against survival time (Fig. 3A). The points that correspond to the events with a censor status of 0 are indicated on the graph. Users can dynamically modify the fold change (up-regulation and down-regulation) thresholds and redraw the plot. A log-rank *P* value is provided as an indication of significance of the difference in survival between any two groups of samples segregated based on gene expression of the gene of interest. The log-rank *P* value is calculated using the Mantel-Haenszel procedure (12). *P* values are recalculated every time a new threshold is selected. Users can toggle to a unified gene expression view with lesser reporters to get a gene-based view of the expression data. To obtain the unified gene expression values, the probe-level data are processed with custom Chip Definition Files that rearrange Affymetrix probes into gene-based probe sets. Probes mapped to alternatively spliced exons are grouped into distinct probe sets. Most 3′ probes are selected for processing. Nonspecific probes are masked before processing. Similar to Kaplan-Meier plots for differential fold change analysis, Kaplan-Meier plots can be drawn for copy number data where genes are mapped to single nucleotide polymorphism probe sets by aligning the probe's physical position to aligned mRNA sequences plus 50 kb upstream and downstream for maximum coverage. Also, Kaplan-Meier plots can be drawn by selecting two patient groups of interest. These groups can be user-defined or predefined lists of patients.

Performing Higher-Order Statistical Analysis on Genomic and Clinical Data Set

Rembrandt supports computer-intensive, high-memory utilizing tasks such as higher-order gene expression analyses (such as class comparison, clustering, and PCA), where the data sets could be as large as 4 GB with an analytic cluster to allow for several simultaneous analytic jobs.

Figure 3B shows an example of a PCA report from the Rembrandt application. This two-dimensional graph plots the various principal components from the gene expression PCA. Various analysis options are provided from which users can select gene/reporter filtering and sample selection settings. Users can click on the three tabs at the top of the graph to display PC1 versus PC2, PC1 versus PC3, or PC2 versus PC3. Each point on the graph represents a sample. The samples are colored by disease type. Users can click on the link on the top left-hand corner of the graph to color by gender. Patients with different survival ranges are indicated by different shapes on the graph. Users can select samples of interest by clicking on the graph and drawing a rectangle around samples to save them for future use.

GenePattern Link

Broad's GenePattern (13) combines a powerful scientific workflow platform with >90 computational and visualization tools for the analysis of genomic data. To expand a researcher's ability to analyze the glioma data sets, Rembrandt has been seamlessly integrated with GenePattern. Shown in Fig. 4A is an expression heat map of 50 additional genes that have expression patterns related to stem cell factor (14) in glioblastoma multiforme.

Plotting Copy Number Data from Patient DNA Samples against Genomic Location

Scatter plots (shown in Fig. 4B) display measured copy number against the physical genome location in an application called webGenome, which has been integrated with Rembrandt. These plots are context sensitive to the copy number reports generated from the copy number queries in the Rembrandt application. Users can view data at arbitrary resolutions from the entire genome on down. When users move the mouse over specific probes, the system provides mouse-over probe names. Clicking on the name of an experiment or bioassay in the plot legend will highlight the corresponding data.

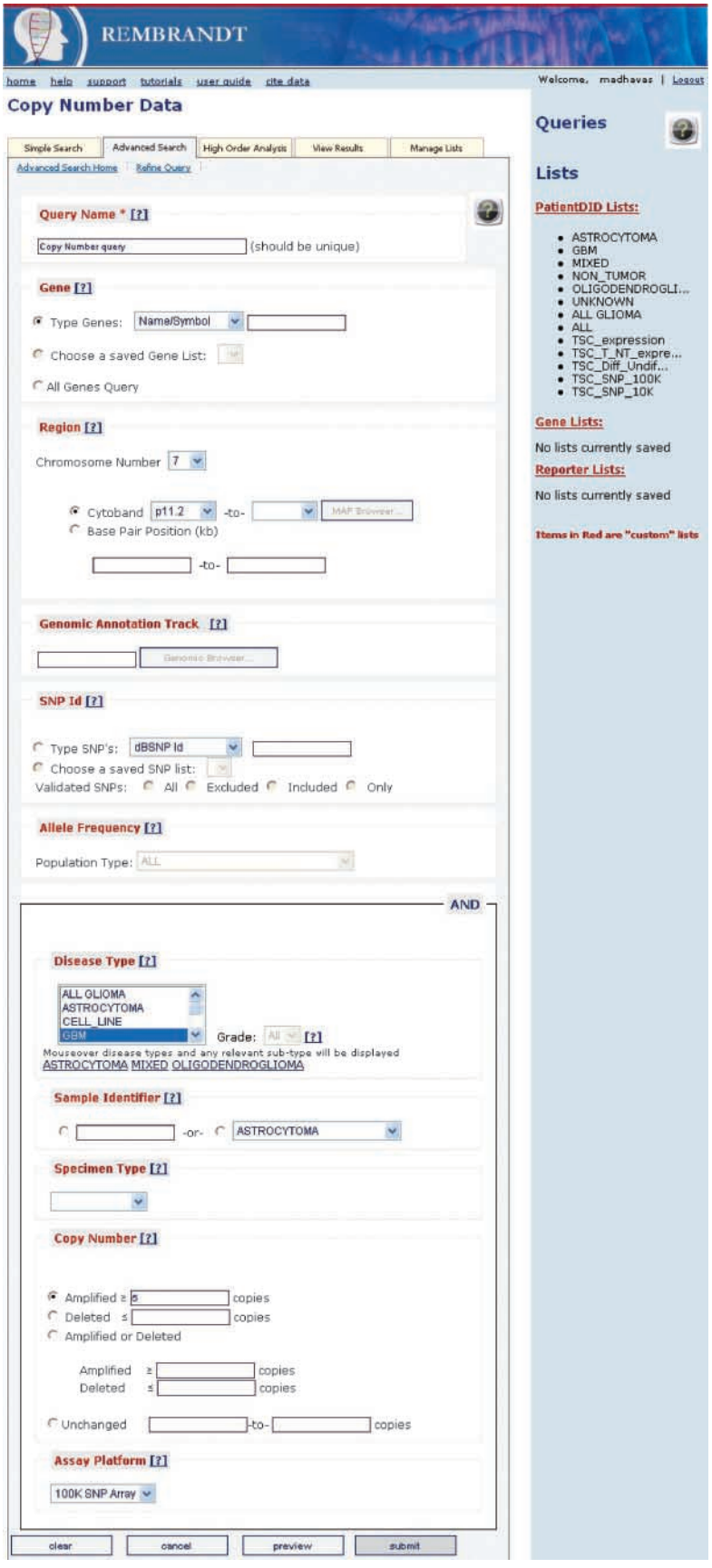
Advanced Query and Report Interfaces

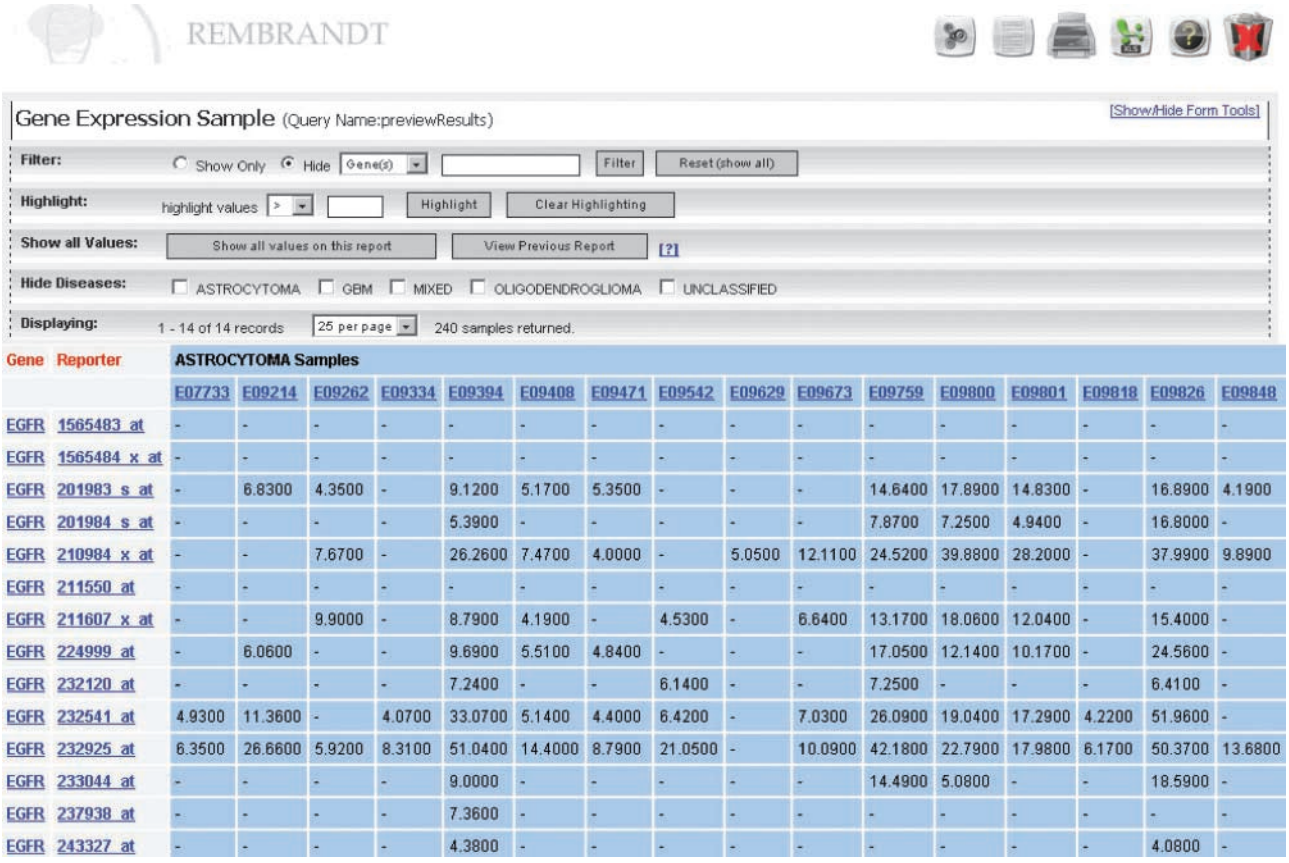
Biomedical researchers struggle to meaningfully integrate their findings across multiple data types. Cancer is a complex disease requiring genomic, proteomic, pathology, imaging, and clinical data for a true understanding of the scope of the problem. Advanced query interfaces (as shown in Fig. 5) in Rembrandt enable this meaningful integration across data types. It allows users to mine the Rembrandt database using various genomic and clinical criteria. These queries can be combined to arrive at reports (shown in Fig. 6) that integrate data from various data domains, such as gene expression, copy number analysis, and clinical trials. Several filtering and data download options are presented in Rembrandt reports.

Rembrandt System Architecture

Rembrandt was developed using a *n*-tier architecture. The system was developed using Java 2 Enterprise Edition, a hybrid star data warehouse schema and various open source technologies. The back end consists of an Oracle 10g database for storing precomputed microarray differential expression, computed copy number, clinical data, and user security information. For performance reasons, normalized gene expression data used by the real-time analysis module are stored as R-binary files. The middle tier, which handles application logic and core functionality, was developed using Java and cancer Biomedical Informatics Grid software development and compatibility guidelines (15). Rembrandt application consists of standard interfaces that enable integration with third-party tools such as caArray, webGenome, and GenePattern. Rembrandt has an Analytical Server that provides on-the-fly computational

FIGURE 4. **A.** Heat-map view in GenePattern. Subsets of data from Rembrandt can be transferred to GenePattern using standard interfaces to invoke several run-time data analysis capabilities. A heat map for 50 neighbors of stem cell factor is shown for astrocytoma and mixed glioma samples in Rembrandt. **B.** Scatter plot for copy number data across physical genomic locations. Scatter plots display measured copy number against physical genome location in an application called webGenome, which has been integrated with Rembrandt via standard data interfaces. These plots are context sensitive to the copy number reports generated from the copy number queries in the caIntegrator application. Users can view data at arbitrary resolutions from the entire genome on down.


REMBRANDT
home help support tutorials user guide site data
Welcome, madhavas | Logout
Copy Number Data
Simple Search
Advanced Search
High Order Analysis
View Results
Manage Lists
Advanced Search Home
Refine Query
Query Name * [?]
Copy Number query
(should be unique)
Gene [?]
Type Genes:
Name/Symbol
Choose a saved Gene List:
All Genes Query
Region [?]
Chromosome Number
7
Cytoband
p11.2
-to-
Base Pair Position (kb)
-to-
Genomic Annotation Track [?]
Genomic Browser...
SNP Id [?]
Type SNP's:
dBSNP Id
Choose a saved SNP list:
Validated SNPs:
All
Excluded
Included
Only
Allele Frequency [?]
Population Type:
ALL
AND
Disease Type [?]
ALL GLIOMA
ASTROCYTOMA
CELL_LINE
GBM
Grade:
[?]
Mouseover disease types and any relevant sub-type will be displayed
ASTROCYTOMA MIXED OLIGODENDROGLIOMA
Sample Identifier [?]
-or-
ASTROCYTOMA
Specimen Type [?]
Copy Number [?]
Amplified ≥ 5 copies
Deleted ≤ copies
Amplified or Deleted
Amplified ≥ copies
Deleted ≤ copies
Unchanged -to- copies
Assay Platform [?]
100K SNP Array
clear
cancel
preview
submit
Queries
Lists
PatientDID Lists:
ASTROCYTOMA
GBM
MIXED
NON_TUMOR
OLIGODENDROGLI...
UNKNOWN
ALL GLIOMA
ALL
TSC_expression
TSC_T_NT_expre...
TSC_Diff_Undif...
TSC_SNP_100K
TSC_SNP_10K
Gene Lists:
No lists currently saved
Reporter Lists:
No lists currently saved
Items in Red are "custom" lists



Gene	Reporter	ASTROCYTOMA Samples															
		E07733	E09214	E09262	E09334	E09394	E09408	E09471	E09542	E09629	E09673	E09759	E09800	E09801	E09818	E09826	E09848
EGFR	1565483_at	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EGFR	1565484_x_at	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EGFR	201983_s_at	-	6.8300	4.3500	-	9.1200	5.1700	5.3500	-	-	-	14.6400	17.8900	14.8300	-	16.8900	4.1900
EGFR	201984_s_at	-	-	-	-	5.3900	-	-	-	-	-	7.8700	7.2500	4.9400	-	16.8000	-
EGFR	210984_x_at	-	-	7.6700	-	26.2600	7.4700	4.0000	-	5.0500	12.1100	24.5200	39.8800	28.2000	-	37.9900	9.8900
EGFR	211550_at	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EGFR	211607_x_at	-	-	9.9000	-	8.7900	4.1900	-	4.5300	-	6.6400	13.1700	18.0600	12.0400	-	15.4000	-
EGFR	224999_at	-	6.0600	-	-	9.6900	5.5100	4.8400	-	-	-	17.0500	12.1400	10.1700	-	24.5600	-
EGFR	232120_at	-	-	-	-	7.2400	-	-	6.1400	-	-	7.2500	-	-	-	6.4100	-
EGFR	232541_at	4.9300	11.3600	-	4.0700	33.0700	5.1400	4.4000	6.4200	-	7.0300	26.0900	19.0400	17.2900	4.2200	51.9600	-
EGFR	232925_at	6.3500	26.6600	5.9200	8.3100	51.0400	14.4000	8.7900	21.0500	-	10.0900	42.1800	22.7900	17.9800	6.1700	50.3700	13.6800
EGFR	233044_at	-	-	-	-	9.0000	-	-	-	-	-	14.4900	5.0800	-	-	18.5900	-
EGFR	237938_at	-	-	-	-	7.3600	-	-	-	-	-	-	-	-	-	-	-
EGFR	243327_at	-	-	-	-	4.3800	-	-	-	-	-	-	-	-	-	4.0800	-

FIGURE 6. Gene expression fold report. All reports in Rembrandt are fully customizable at the report window, making it unnecessary to re-run queries to refine the results.

analysis capability. The Analytical Server communicates asynchronously with Rembrandt's middle tier via the Java Messaging Service. Java Messaging Service allows Rembrandt to abstract the statistical packages being used for heavy computational tasks.

Rembrandt Cancer Biomedical Informatics Grid Service

Basic and clinical research has increasingly become dependent on advanced information technologies for management, exchange, and analysis of diverse biomedical data. Although a wealth of information is collected by the cancer research community, any one given researcher is faced with challenges in discovering, extracting, and analyzing the information relevant to his/her research. To address this need, the National Cancer Institute has initiated a national-scale effort, called the cancer Biomedical Informatics Grid, to develop a federation of interoperable research information systems. At the heart of the cancer Biomedical Informatics Grid approach to federated interoperability effort is a Grid middleware infrastructure, called caGrid (16). caGrid Data Services provide the means to share data via the caGrid federated infrastructure. One of the major goals of the current release of Rembrandt was to create a clinical genomic object model and expose the domain model through a caGrid data service. The purpose of the object model is to help capture the relationships between the clinical study and its associated experimental observations. The Rembrandt caGrid service can be used to obtain programmatic access to public data in Rembrandt in a federated fashion and can be found at http://caintegrator.nci.nih.gov/wsrf-rbt/services/cagrid/RembrandtGridService.

Conclusion

Large-scale data sets from genomics, proteomics, population genetics, and imaging are driving research at a previously unprecedented pace. Bioinformatics data management providers must serve these data sets in a usable way that helps find the needle in a haystack effectively and accurately. The goal of the "omic" sciences is not to generate numbers but rather "insight." The Web interfaces are burdened with displaying terabytes of data in ways that physician scientists can comprehend and use the results to develop hypothesis for

FIGURE 5. User-friendly data query interface. Query pages enable users to restrict their searches in the database to specific genomic and/or clinical criteria.

their next study or trial. Ultimately, we feel that information must be standardized, integrated, and made available at the point of care to help patients and physicians make optimal decisions.

Tools such as Rembrandt have primarily focused on the usability aspect of high-throughput heterogeneous data and yet enabling power users and bioinformaticians to tap into runtime analysis tools such as gene pattern or use the programmatic interfaces that are provided via the caGrid service. From a technical standpoint, the Rembrandt platform provides developer tools for a highly scalable system to include new data types (as shown in Fig. 1) and connect with existing ones to present integrated data views to users. This flexible discovery informatics platform has aided in implementing data portals to host several other cancer clinical data sets including those from the I-SPY stage III breast cancer study and The Cancer Genome Atlas (TCGA; ref. 17) project data included in the Cancer Molecular Analysis Portal. In this respect, it is worth to point out that the new Cancer Molecular Data portal has reutilized many of the features available in Rembrandt to suit a more general set of tumor sample analysis. At the sample level, GMDI and TCGA are complementary in many levels. GMDI is a prospective study wherein 14 institutions recruited patients with any type of glioma giving a wide spectrum of demographic sampling due to the geographic dispersion of the sites. The TCGA sample collection pipeline included two centers that had retrospective sample collections of glioblastoma multiforme. Thus, TCGA focused its analysis on high-grade glioblastoma multiformes, whereas the samples in GMDI represent all glioma grades and subtypes described in the WHO classification, allowing for studies on the differences of gliomas as they progress. The clinical data obtained by the GMDI project are comprehensive, because the study was conceived as a prospective, natural history clinical trial, thus allowing for the collection of a wide range of clinical data points. On the other hand, the TCGA project, in virtue of its more focused nature, has produced more molecular data types (methylation, sequencing, and miRNA expression) than GMDI. However, the GMDI samples are being used to acquire those data types, and they will be incorporated to Rembrandt as sufficient numbers of samples are processed.

The ultimate beneficiaries of Rembrandt are the brain tumor patients themselves. Rembrandt is designed to bridge the gap between biological and clinical information to help patients receive a better, biologically oriented therapy tailored to their specific needs. As such, we plan to incorporate new and useful capabilities in future releases that are not available at present time, such as the ability for researchers to incorporate their own data to the system to compare with the large data set already in the database. It is hoped that the GMDI and Rembrandt will provide a much needed resource for scientists and physicians combating brain cancer, and ultimately other forms of cancer, for providing the data and bioinformatics tool set that may allow the development of a biologically and clinically significant pathologic classification of brain tumors and help elucidate novel molecular targets for therapy.

Availability

Rembrandt is freely available to all users at https://caintergator.nci.nih.gov/rembrandt. The source code for Rembrandt is also available under a nonviral cancer Biomedical Informatics Grid license at https://gforge.nci.nih.gov/frs/download.php/1489/rembrandt_1_0.zip. The Rembrandt caGrid service is accessible at http://caintegrator.nci.nih.gov/wsrf-rbt/services/cagrid/RembrandtGridService.

Web Resources

Rembrandt clinical genomics object model: http://Rembrandt.nci.nih.gov/content/Rembrandtlfs/RembrandtEA1.0docs/index.htm.

Rembrandt clinical genomics data model: http://Rembrandt.nci.nih.gov/developers/images/db_model2.jpg.

Rembrandt application: http://rembrandt-db.nci.nih.gov.

Rembrandt information site: http://rembrandt.nci.nih.gov.

webGenome: http://webgenome.nci.nih.gov/webgenome/home.do.

GenePattern: http://www.broad.mit.edu/cancer/software/genepattern/.

caArray: https://array.nci.nih.gov/caarray/home.action.

I-SPY trial: http://ncicb.nci.nih.gov/tools/translation_research/ispy.

TCGA: http://cancergenome.nih.gov/.

Cancer molecular analysis portal (access to TCGA data sets): http://cma.nci.nih.gov.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Anand Basu, Shine Jacobs, Alex Jiang, Huaitian Liu, Ram Bhattaru, Michael Harris, Kevin Rosso, Ryan Landy, Hangjiong Chen, and Ying Long for contributions to Rembrandt software development and data loading; George Komatsoulis for reviewing data sharing policies and contributing to the Rembrandt domain information model; Carl Schaefer and Tracy Lively for reviewing usecases and interim releases of the software; Juli Klemm for helping with the integration of Rembrandt with caArray data repository; Jill Hadfield for technical documentation; David Hall, Dean Jackman, and Vessalina Bakalov for efforts on webGenome; and The University of Texas M. D. Anderson Cancer Center Data Management Initiative team for making the clinical reports available from the NABTC GMDI study for populating the Rembrandt database.

References

1. Cancer Statistics Branch, NIH. Cancer survival rates. In: Harras A, editor. Cancer: rates & risks. Washington (DC): U.S. Department of Health & Human Services, NIH; 1996. p. 28–34.

2. Nutt CL, Mani DR, Betensky RA, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res 2003;63:1602–7.

3. Mischel PS, Shai R, Shi T, et al. Identification of molecular subtypes of glioblastoma by gene expression profiling. Oncogene 2003;22:2361–73.

4. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 2006;9:157–73.

5. Nigro JM, Misra A, Zhang L, et al. Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. Cancer Res 2005;65:1678–86.

6. Louis DN, Ohgaki H, Wiestler OD, et al. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol (Berl) 2007;114:97–109.

7. Miller CL, Diglisic S, Leister F, Webster M, Yolken RH. Evaluating RNA status for RT-PCR in extracts of postmortem human brain tissue. Biotechniques 2004;36:628–33.

8. Matsuzaki H, Dong S, Loi H, et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 2004;1:109–11.

9. Lee J, Kotliarova S, Kotliarov Y, et al. Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. Cancer Cell 2006; 9:391–403.

10. Galli R, Binda E, Orfanelli U, et al. Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. Cancer Res 2004;64:7011–21.

11. Lee J, Son MJ, Woolard K, et al. Epigenetic-mediated dysfunction of the bone morphogenetic protein pathway inhibits differentiation of glioblastoma-initiating cells. Cancer Cell 2008;13:69–80.

12. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.

13. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006;38:500–1.

14. Sun L, Hui AM, Su Q, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell 2006;9:287–300.

15. Phillips J, Chilukuri R, Fragoso G, Warzel D, Covitz PA. The caCORE Software Development Kit: streamlining construction of interoperable biomedical information services. BMC Med Inform Decis Mak 2006;6:2.

16. Oster S, Langella S, Hastings S, et al. caGrid 1.0: an enterprise Grid infrastructure for biomedical research. J Am Med Inform Assoc 2008;15: 138–49.

17. McLendon R, Friedman A, Bigner D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.